软件开源研究的主题识别及演化分析

董平军 高翔菲
东华大学旭日工商管理学院 上海 200051

摘要：［目的/意义］软件开源是社会化软件生产中一种重要的生产组织方式和协同创新运动。通过对国内外软件开源相关研究的主题识别及演化分析，探究软件开源研究领域的阶段性热点和趋势变化规律，为以促进中国软件开源创新进一步优化发展为主旨的学者开展研究梳理方向。［方法/过程］以从 Web of Science 数据库检索到的 2001 年至 2023 年 5 月 10 日期间的软件开源领域文献作为语料库，采用困惑度指标确定主题数目，训练 LDA 主题识别模型得到主题-词分布和文档-主题分布，根据主题-词分布对主题进行标识，依据文档-主题分布计算主题强度，进而识别热点主题和归纳演化路径。［结果/结论］主题识别结果表明，软件开源研究领域存在六个重要主题，分别是贡献动机、商业模式、开源治理、协作模式、开源协议、企业参与；从主题演化角度上看，软件开源在商业模式、开源治理和企业参与主题上近年来具有相对较高的研究热度，开源协议的研究趋势相对稳定，贡献动机和协作模式的研究热度虽然呈相对下降趋势，但自始至终一直保持较高的受关注度。软件开源研究呈现由关注开源动机自发、自治的个人维度到企业、政府参与的组织维度的发展规律。建议学者们关注中国情景下开源生态各类主题研究，为我国开源生态健康发展提供理论支持。

关键词：软件开源 主题识别 主题演化 LDA 模型
分类号：G353.1

1 引言

近年来，软件开源作为社会化软件生产中一种重要的生产模式和协同创新运动，形成了丰富的开源生态，正在产生越来越大的影响。在软件开源生产模式下，开源生态的软件使用者、软件开发者、开源组织者、开源系统平台、开源协议、开源研究者等要素互相营养、相互协同、持续迭代：软件使用者自由下载试用软件并反馈信息，软件开发个人或组织可以在开源许可条款下查看、创新并再分发源代码，开源组织通过开源协议和章程负责维护整个开源生态的发展等。软件开源是共享经济在软件生产领域的体现，与软件闭源生产模式一起组成整个软件生产的生态大系统。

在过去的几十年中，软件开源社会化运动持续推动软件生产发展，不断引领创新潮流。从互联网早期时代的 Linux 操作系统、MySQL 关系型数据库管理系统，到移动时代的安卓操作系统、Web 服务器 Apache、前端开发框架 VUE，软件开发工具如 Eclipse、vsCode，项目管理协同工具 git 等等，再到如今引爆时代的 Python 语言和 ChatGPT，都是受益于开源模式下的创新产物。开源运动形成了以 github 为典型代表的开源平台社区，开源创新潮流已经不仅仅局限于软件生产，开始触达硬件、文档、音乐等更广泛的领域。在开源生产模式下，全球范围内各种专业知识技能和生产要素密切融合、快速迭代，从而释放出创新的巨大潜能，推动技术与社会的不断进步。

基于开源模式在软件创新生产中的积极作用和我国软件发展相对滞后的局面，我国自 2009 年开始积极构建和发展开源生态系统。截止到 2022 年我国开发者数量增长全球排名第一，根据国内头部开发者社区 CSDN 的统计，我国开发者用户注册超 3500 万，其中超过 94%的开发者正在使用开源，超过 40%的开发者参与过开源项目的建设，根据开源社区 Gitee 的统计，2021 年 Gitee 新增注册用户超过了 180 万，累计开源开发者超过 800 万；根据 Github 2021 年数据统计，我国开发者数量已经增长至 755 万，全球排名第二。2022 年在国际顶级开源基金会中，中国开源表现出强烈的贡献积极性和参与热情，在开源基础设施基金会中，我国董事会成员占比超过 40%，在云原生计算基金会中来自我国的开源项目超过 20%，在 Apache 软件基金会中，来自我国的

作者简介：董平军(ORCID：https://orcid.org/0000-0002-2889-6289)，硕士生导师，副教授，博士；高翔菲(ORCID：https://orcid.org/0009-0006-9450-8041)，硕士研究生，E-mail：2114085146@qq.com。

活跃开源项目有 24 个，其中 14 个是顶级开源项目，特别是 2021 年，全球仅有 5 个开源项目进入 Apache 软件基金会孵化器，这些新项目全部来自我国。此外，越来越多的中国企业意识到开源的重要性，积极参与到开源项目中。一些互联网巨头公司，如华为、阿里巴巴、百度等，积极开源自己的软件和技术，并投资于开源生态系统的发展，推动开源项目的成长。政府也对开源发展提出了要求，2016 年 12 月 18 日，工信部印发的《大数据产业发展规划（2016-2020 年）》中明确提出要“鼓励开发者、企业、研究机构积极参与大数据开源项目，增强我国在开源社区中的影响力”[1]；2021 年 3 月 12 日，开源首次写入《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》[2]，明确提出支持数字技术开源社区等创新联合体发展；2021 年 11 月 30 日，工业和信息化部印发《“十四五”软件和信息技术服务业发展规划》，系统布局“十四五”开源生态发展[3]；2022 年 01 月 12 日，国务院印发《“十四五”数字经济发展规划》，提出支持具有自主核心技术的开源社区、开源平台、开源项目发展，推动创新资源共建共享，促进创新模式开放化演进[4]。

综上所述，我国的开源创新、开源运营等正处于加速阶段，软件开源生态系统得到了较好地发展，但作为开源生态系统组成部分之一的软件开源研究相对较少。

本文拟通过对国际上软件开源领域研究的进行主题识别及主题演化分析，从而能够了解开源软件生态的发展规律，为我国学者开展中国情景下的软件开源共享研究提供有益的启示和指导，从而最终推动我国软件开源共享创新快速发展，同时为国际上开源共享发展贡献中国力量。

2 主题识别及演化相关研究

主题代表着文本中特定的信息内容或关注点，是指文本中的一种概念或话题的集合，可以由一组相关的词汇或短语抽象表示，能够帮助读者快速了解文本的核心内容，从而提供更好的文本摘要和信息概览。主题识别是一种从给定文本集合中发现隐藏在其中主题的文本挖掘技术，它通过分析文本的语义和上下文信息，推断出文本中的主题及相应的概率。主题演化是指在一定时间范围内，主题在不同时间点上的变化和发展过程。主题识别及演化分析通过运用文献计量学或自然语言处理等方法，对一个领域中的主题进行识别和跟踪，并对其发展趋势进行动态分析和可视化呈现，其研究内容通常涉及主题的演化路径、关键词共现网络的形成和演化、领域内核心作者和机构等方面的变化，以及影响主题演化的因素和机制等。主题识别及演化分析在多个领域都有应用，如舆情监测[5]、社交媒体分析[6]、医疗健康[7]、商业智能[8]和政府决策[9]等，主题识别及演化分析对研究者更加全面地了解不同领域在不同时期的研究进展和变化趋势有重要意义。

主题识别及演化分析常用方法技术的优缺点如表 1 所示。文献计量学方法通常基于词频、共现关系以及共引分析识别文档的主题[10]，是一种简单易用的主题识别方法，有很多学者基于文献计量学方法进行主题识别和演化分析[11][12][13]，其中，社会网络分析在揭示主题之间的关系方面备受青睐，邢晓昭等[14]以文本挖掘和社会网络分析为方法手段，提出基于主题演化的颠覆性技术识别方法；Reza Vahidzadeh 等[15]基于社会网络分析探究区域产业共生(RIS)研究领域中技术和非技术两个方面的主要主题和趋势。基于文献计量学的方法进行主题识别尽管可以挖掘大规模文献的主题，但容易忽视文本信息中的语义信息，导致结果缺乏丰富性，因此，越来越多的学者使用基于机器学习的主题模型方法进行主题挖掘和演化分析[16][17][18]，采用机器学习算法，如朴素贝叶斯、支持向量机等，对文本数据进行分类和聚类分析，可以实现主题的识别和演化的跟踪。在基于机器学习的主题识别和演化分析中，利用概率图模型对文本数据进行主题建模被广泛使用。常用的概率图模型如潜在语义分析（Latent Semantic Analysis，LSA）、潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）和动态主题模型（Dynamic Topic Model，DTM）等将文本数据转化为主题-词语分布的表示形式，将文档、主题和词汇之间的关系进行隐含的建模，通过学习概率分布参数来发现文本数据中的主题结构和演化模式，能够挖掘出文本中的潜在语义关系，在主题识别和演化分析中比传统方法更有效[19]。

表 1 主题识别及演化分析常用方法比较

方法	分类	优点	缺点
传统的主题识别方法	词频分析； 词语共现分析； 社会网络分析； 因子分析；	1. 相对简单直观，易于理解和解释； 2. 能够揭示出文献之间的引用关系和影响力；获得作者及国家之间的合作关系。	1. 关键词选取具有较强的主观性，通常会抽取高频的关键词； 2. 忽略了文本内部的语义信息，只关注文献之间的关系和引用情况。
基于词向量的主题识别方法	动态主题模型； 潜在语义分析； 潜在狄利克雷分布； 动态主题模型； 聚类分析	1. 采用概率图模型，为每个主题提供词语的概率分布，揭示主题内部潜在的丰富语义内容； 2. 可以处理大量的数据，更加快速、准确；主题抽取的主观性大大降低。	1. 需要大量的训练数据来进行模型构建和训练； 2. 在处理大规模文本时，计算复杂度较高；

综上所述，已经有很多学者在软件开源领域进行了大量的研究，但是现有的文献分析存在一些局限性。首先，一些研究只关注单一的开源软件项目或应用领域，分析软件开源领域的特定问题，而对于整个软件开源研究主题的分析尚不充分。其次，一些研究综述采用传统的基于文献计量学的方法[20]，从无监督学习的主题建模视角对开源软件进行主题识别的研究较少，LDA 模型尚未在开源软件领域中得到广泛应用，缺乏对主题演化的深入探究。此外，个别研究仅基于开源平台的源代码存储库中的数据进行分析，识别开发人员贡献的主题，以便更好的为开发人员匹配合适的任务[21]，没有充分利用其他数据来源识别开源领域的研究主题。因此，本文旨在使用 LDA 主题建模技术，对软件开源领域的主题和演化进行深入研究，探究软件开源领域的发展趋势和未来方向。本文的研究能够为软件开源的发展提供参考和建议，在一定程度上促进开源共享领域的研究与发展，并对于其他领域的主题分析和演化研究提供借鉴和参考，具有一定的现实意义。

3 研究设计

3.1 数据采集与获取

本文选择 Web of Science 数据库（以下简称 WOS）作为数据的来源，检索时间为 2023 年 5 月 11 日，检索文献的时间范围为 2001 年至当前检索时间，文献类型分别选择“Article”，以排除会议、报纸等文献。以 TS=("open source software" OR "free software" OR "libre software")作为主题词进行检索，初步筛选得到论文 1146 篇，通过逐一阅读论文标题和摘要部分，了解论文主题，排除研究方向开源软件应用与技术问题等与本论文研究无关的文献，最终得到 769 篇文献的题录信息，题录信息包括题目、关键词、关键词扩展、摘要、期刊来源、出版年份、作者、机构、国家等信息。每年的文献数量分布如图 1 所示。

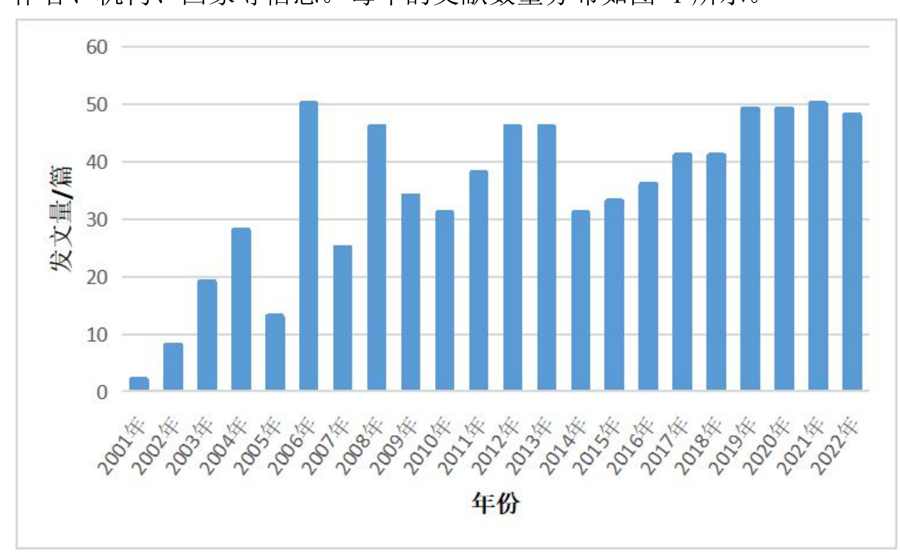


图 1 文献数量分布图

图 2 显示了文献数量排名在前十五位的来源期刊，可以看出，来源于《RESEARCH POLICY》、《JOURNAL OF SYSTEMS AND SOFTWARE》、《INFORMATION SYSTEMS RESEARCH》的文献数量相对较多，在排名前十五位的来源期刊中文献数量占比 30.7%。

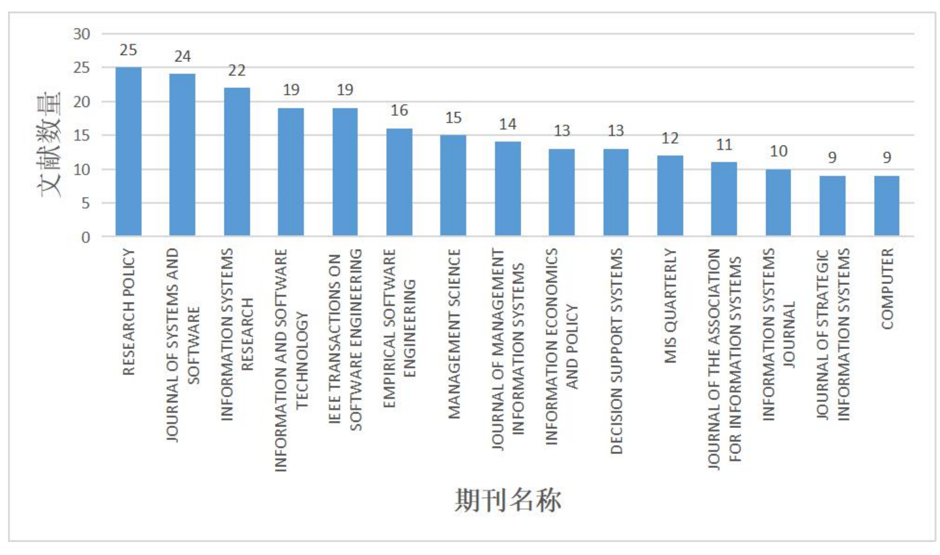


图 2 主要来源期刊

3.2 数据预处理

在进行主题识别和演化分析之前，需要对收集的文献数据进行预处理，以提高数据质量和准确性。本文提取从 WOS 中导出的文献题录信息中的标题、关键词、关键词扩展、摘要作为主题模型的语料来源，调用 NLTK 自然语言处理工具包在进行分词处理，主要包含以下处理：①特殊的字符被排除在外，如数字、标点符号等；②将缩写还原成完整形式，以合并同一概念的不同形式，如“OSS”还原为“Open Source Software”，“PR”还原为“Pull Request”等；③将单词转换为小写形式，标记词性并进行词性过滤，去除形容词、副词等没有意义的单词；④构建停用词表、同义词表和开源软件研究的专有名词词典，完成停用词去除、合并同义词的操作，并在训练模型过程中持续扩充停用词表、同义词表和专有名词词典，不断优化分词结果。⑤将单词进行词形还原和词干化处理。

3.3 研究方法

本文采用基于概率图模型的常用的主题建模技术 LDA（Latent Dirichlet Allocation）模型对软件开源的主题特征进行分析，揭示软件开源领域的关键主题和研究热点。Blei 等人[22]在 2003 年提出了 LDA 主题建模方法，通过对文本中的单词进行主题分布和主题中单词分布的联合建模，识别文本中的潜在主题。LDA 模型可以将软件开源领域的大量文本数据进行主题建模，识别出其中的关键主题和互相关联的词汇，挖掘出研究关注的核心议题。它的优势在于可以从文本中自动发现主题，不需要预先定义主题或手动标注样本。它能通过对文本数据的统计分析，自动学习主题和文档之间的关系。LDA 模型在处理大规模文本数据时非常有用，并且可以发现不同时间阶段的主题演化，需要预先设定主题个数等信息[23]。LDA 是一个三层贝叶斯模型，如图 3 所示，它用主题上的概率分布表示每个文档，其中每个主题都表示为单词上的概率分布。LDA 模型假设一篇软件开源研究文章的每个词是通过以一定概率选择某个主题，并从该主题中以一定概率选择某个词语的过程得到的，其中文档的主题分布和主题的单词分布分别取决于由α和β参数决定的 Dirichlet 先验分布。

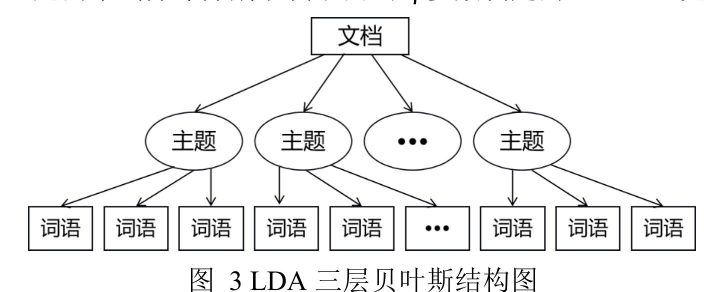


图 3 LDA 三层贝叶斯结构图

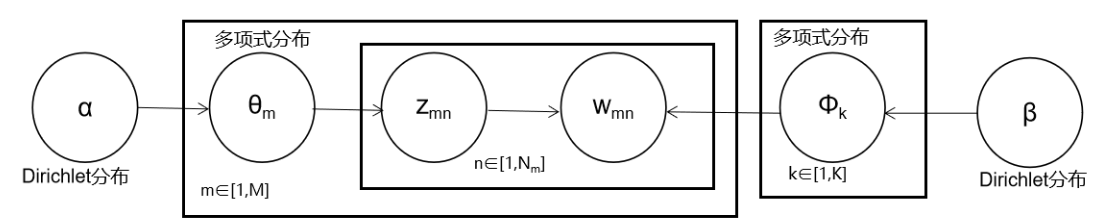


图 4 LDA 模型图

LDA 模型现在已被广泛用于发现文档集合中的潜在主题，其原理如图 4 所示。本文将一篇软件开源研究文献的标题、关键词、关键词扩展、摘要作为一篇文档，LDA 模型假设在 M（769）篇关于软件开源的文档中，每篇文档中由 K 个主题组成，每个主题下的 N 个词语w_{mn}构成这篇文档；每篇文档在主题上服从多项式分布，每个主题在单词上服从多项式分布；每篇文档的主题的多项式分布的先验分布是参数为α的狄利克雷分布，每个主题的词汇的多项式分布的先验分布是参数为β的狄利克雷分布。

基于以上假设，对于 769 篇软件开源研究领域中的每篇文档，LDA 模型生成文本的基本流程是：①从参数为α的狄利克雷分布中采样，随机生成第 m 篇文档对应主题的多项式分布θ_m；②根据多项式分布θ_m随机生成第 m 篇文档中第 n 个单词的主题z_{mn}；③从参数为β的狄利克雷分布中采样，随机生成主题z_{mn}对应单词的多项式分布φ_k；④综合主题z_{mn}和主题z_{mn}对应的单词的分布情况φ_k生成单词w_{mn}；循环上述过程生成一个包含 N 个词语的文档，最终生成 K 个主题下的 M（769）篇文档。采用 Gibbs 采样方法进行参数估计，就可以训练出每篇文档的主题分布θ_m及其对应的单词分布概率φ_k。

3.4 LDA 参数确定

Gensim 是一个用于文本分析、主题建模和词嵌入等自然语言处理任务的 Python 软件包，本文利用 Gensim 包训练 LDA 模型。在训练 LDA 模型之前，首先需要确定 3 个超参数α、β、k，超参数α代表每篇文档下主题的狄利克雷分布先验参数，控制文档-主题分布的稀疏性以及每个文档中主题的多样性程度。α值越小，每个文档包含的主题越少。β代表每个主题下词汇的狄利克雷分布先验参数，定义了每个主题中词语的多样性程度。较小的β值会使每个主题包含少数几个高频词语，较大的β值会使每个主题更均匀地包含各个词语。由于主题结果对α、β参数的值不是很敏感[24]，α、β一般选取 Gensim 包中 LDA 模型的默认值，即使用固定的对称先验 1/k，k 代表最优主题的数量。

对于最优主题数 k 的确定，常用的方法有四种：①根据困惑度确定[25]：困惑度通常用于衡量主题模型模型预测新数据的准确性。通常困惑度越小，模型的预测性能越好。②根据一致性分数确定[26]：基于一致性度量来评估主题的质量。通过计算主题中各个词的一致性得分确定最优主题数目，一致性分数越高，主题模型越好。③使用文本聚类方法确定主题提取数目[27]：通过对文本进行聚类，从聚类结果中获取 LDA 主题模型的数量。这种方法容易受到样本本身特点和聚类算法的影响，并且需要人工干预，往往通过多次试错才能找到合适的主题提取数目。④根据主题分布可视化及人工评估确定[28]：通过可视化来辅助确定主题数。通常采用主题-词汇分布图来表示主题，通过观察不同主题数目下的图像特征，阅读相关领域权威文献，结合研究目的和领域知识来进行判断，比较不同主题数量下的主题质量、主题之间的相关性和可解释性，选择一个最优的主题数量。

困惑度表示文档 d 从属的主题的不确定性[29]，用于评估语言模型的性能，是当前研究中确定主题数目最受欢迎的方法，因此，本文采用困惑度指标来确定主题数目的大小，当困惑度达到最小或处于转折点处时，主

题模型的泛化能力强，此时得到最优主题数量 k，困惑度计算公式如公式 1 所示：

$$\mathrm{Perplexity}(D) = \exp\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\} \quad 公式 1$$

$$p(w_d) = \prod_{i}^{N_d} \sum_{z} p(w_{d,i}|z)\, p(z|d) \quad 公式 2$$

其中，M 是软件开源领域的文档总数，$p(w_d)$为文档 d 的生成概率，N_d 为文档 d 包含的单词总数。$p(w_d)$的计算方式为其文档中每个词汇生成概率之积，如公式 2 所示。

将主题数量限制在 2-20 之间，依次计算各个主题数量的困惑度，绘制困惑度随主题数量变化的折线图。从图 5 中可以看出，当主题数量为 6 时困惑度最小，用一致性分数进行验证，图 6 显示主题数量为 3 时一致性分数达到第一个转折点，但此时主题数量过少，模型泛化能力弱，在主题数量为 6 时达到第二个转折点，此时一致性分数也比较高，因此本文选定开源软件研究领域的主题数量为 6。

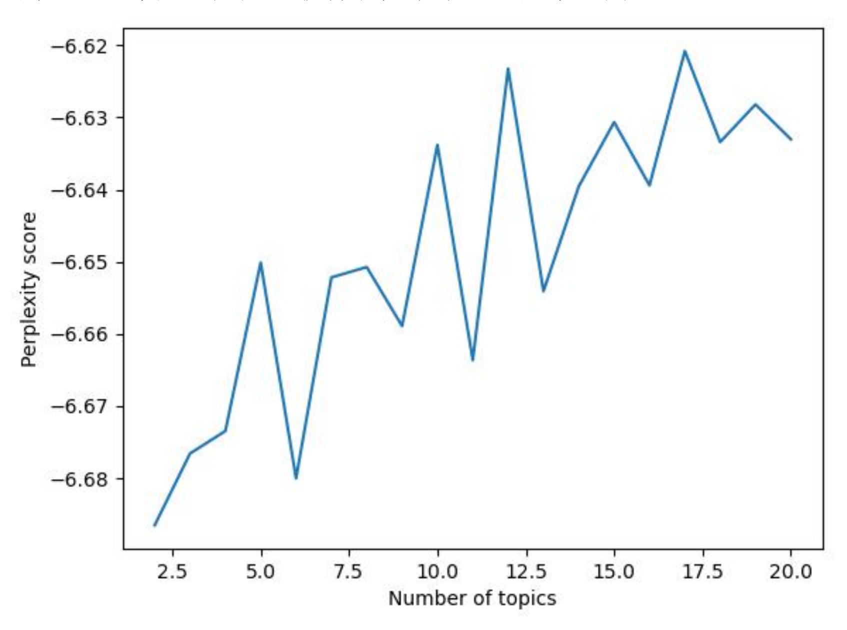


图 5 主题数量-困惑度变化情况

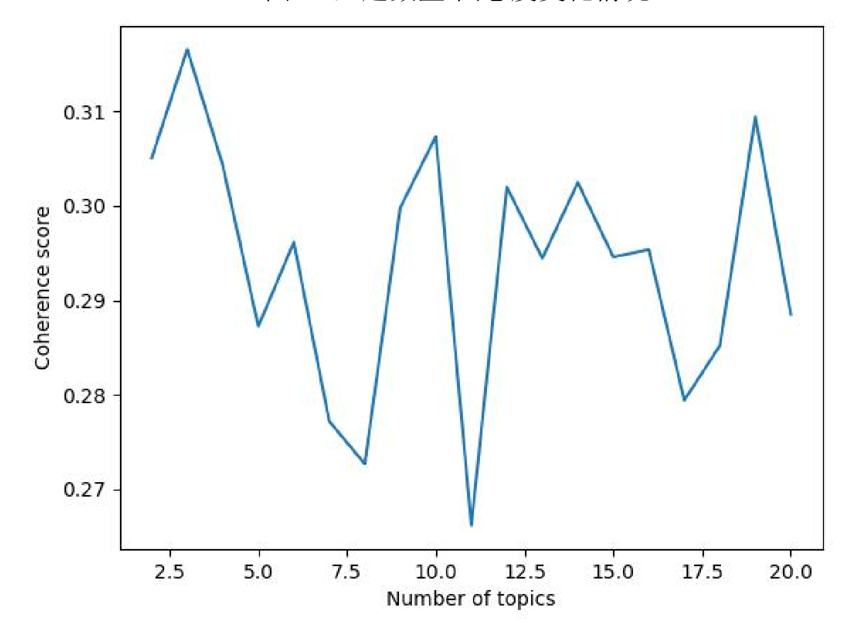


图 6 主题数量-一致性分数变化情况

3.5 LDA 主题建模

设置主题数量为 6 并训练 LDA 模型，得到“主题-词”分布和“文档-主题”分布两个结果。“主题-词”分布表明每个主题包含哪些单词及这些单词在该主题中的概率，可以用来研究文本中不同主题之间的关系、找出文本中特定主题的内容以及揭示主题内容的演化路径。“文档-主题”分布表明每篇文档包含哪些主题以及这些主题在该文档中的权重，主要用于主题强度计算、识别阶段性热点主题以及判断主题演化路径。

本文基于 Sievert 和 Shirlet 等人[30]在 2014 年提出的 Web 的主题可视化方法，利用 LDAvis 进行交互式可视化，以更好地理解和分析 LDA 模型中的主题、单词及其权重分布，从整体的角度来观察主题和主题以及主题和词语之间的关系。可视化结果如图 7 所示，图中 6 个大小不同的气泡代表六个主题，它的大小与包含的文档数量有关，表示该主题在整个文档集中的相对重要性。主题之间的距离反映了它们之间的相似度，距离越近的主题具有更相似的单词分布，图中各主题之间交叉很少，说明分类效果较好。

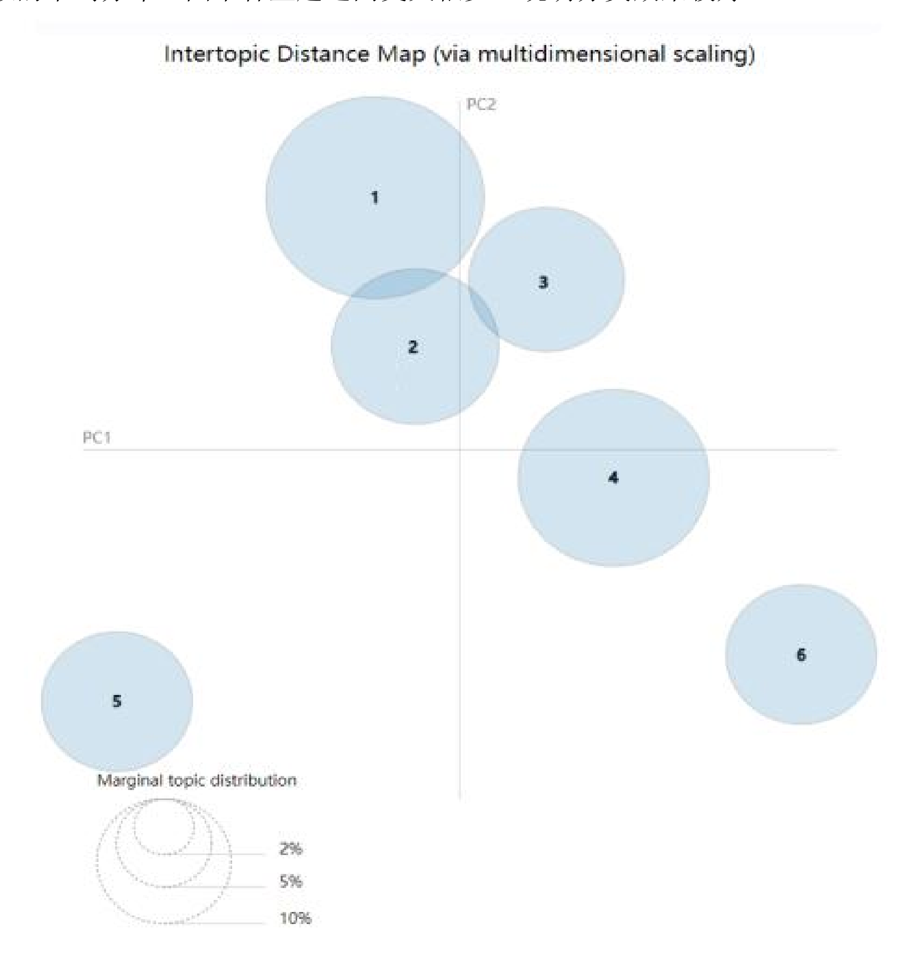


图 7 主题识别可视化结果

Figure 7 Topic Recognition Visualization Results

4 数据结果

4.1 主题归纳分析

根据 LDA 模型得出的“主题-词”分布对主题进行标识，选取每个主题下概率最高的前十个词汇判断主题内容，最终主题标识的结果如表 2 所示。

表 2 LDA 主题建模结果

Topic1（贡献动机）		Topic2（商业模式）		Topic3（开源治理）	
project	0.038710095	development	0.020549098	project	0.027586445

motivation	0.037083324	community	0.020520825	governance	0.012733819
community	0.022500135	project	0.015811738	development	0.012479639
developer	0.016605742	firm	0.014142942	adoption	0.011333821
contribution	0.016531823	model	0.013599659	base	0.010658319
study	0.014288876	innovation	0.013513555	study	0.010217006
development	0.014029684	business	0.008755206	model	0.010209058
innovation	0.013139865	study	0.008111687	network	0.009370943
work	0.012944275	process	0.008107803	ecosystem	0.009038342
research	0.010772946	develop	0.008089474	effect	0.008712007

表 2 （续表）

Topic4（协作模式）		Topic5（开源协议）		Topic6（企业参与）	
project	0.04877134	license	0.031450167	community	0.036629133
developer	0.034251563	study	0.018310972	project	0.031768467
development	0.022790086	innovation	0.016253594	firm	0.017995713
network	0.019850846	process	0.015639516	developer	0.015662977
community	0.017864436	firm	0.013896711	study	0.011578429
knowledge	0.016069325	model	0.011127581	design	0.010776361
model	0.013099793	case	0.010097274	model	0.009528774
team	0.011104287	ecosystem	0.009492678	development	0.008961064
study	0.010424438	system	0.009243087	research	0.008455344
impact	0.009592903	research	0.009092568	analysis	0.007370888

依据表 2 每个主题下的高概率单词，总结归纳最符合当前主题下高概率单词的主题名称。在主题 1 中，项目、动机、开发者、贡献等高频单词与开发者参与项目的动机研究比较贴切，被标注为“贡献动机”。贡献动机主题是研究者最早关注的话题之一，早期很多学者针对 Lerner 和 Tirole 提出的问题“为什么成千上万的顶尖软件开发人员会免费为创建公共产品做出贡献？”[31]展开研究。Alexander Hars 和 Shaosong Ou 采用电子邮件问卷调查的方法系统地研究了为什么人们会参与开源软件项目，揭示了参与动机的内在因素和外在因素两个方面[32]，学习也是激励人们参与 OSS 社区的主要驱动力之一[33]。von Hippel 和 von Krogh 从组织科学的角度探究开源软件参与者的动机，提出了一种“私人-集体”的创新激励模式[34]。Shaul Oreg 和 Oded Nov 根据工具性的高、中、低三个等级将内在动机和外在动机进一步分成三类[35]。von Krogh 等在 2012 年构建了一个动机-实践的理论框架，扩展了对个人动机的假设，将短期奖励之外的长期、有价值的追求纳入其中[36]。Moqri, M 等证实了开发人员参与开源软件开发不仅存在非货币激励，还存在与未来货币奖励相关的激励[37]。

主题 4 中的开发者、团队、网络、影响等高频词与开发者的协作模式有关，被标注为“协作模式”。协作模式主题也是学者较早研究的主题，Raymond 的《大教堂与集市》系统性地解释了开源软件开发者的协作模式，他提出以 Linux 为代表的开源软件使用了集市的开发模式，大多数传统的商业软件则采用大教堂的开发模式[38]，并且与大多数商业软件相比，开源软件项目的开发人员以更自由的工作方式组织和贡献项目，他们之间的协作互动产生了一个不断演化的的开发者社交网络[39]，当一个开发人员与项目发起人之前有很强的协作关系时，他更有可能加入一个项目[40]。在开源项目中，开发者之间的协作模式起着关键的作用，决定着项目的成功和成长。

主题 5 中许可证、创新、成功等高频词揭示了对开源软件项目许可证的选择以及项目成功的研究，被标注为“开源协议”。许可证的选择对开源项目的成功与可持续发展十分重要，许可证选择与组织赞助交叉作用影响贡献者对开源软件开发项目的兴趣和开发活动，进而影响开源项目的成功[41]。此外，开源项目的成功还与

开发人员的持续参与[42][43][44]、开源软件社区的文化和意识形态[45][46][47]等密切相关。管理开源软件开发中的许可证遵从性是当今的一个重要问题，不遵守许可证协议会导致组织声誉的损失[48]，可以根据开发衍生软件需要付出的努力选择许可证，当开发软件需要付出大量的努力时，较少限制的许可证更有利于项目的成功[49]。目前比较流行的几种开源许可证包括 MIT 许可证、GNU 通用公共许可证（GPL）、 Apache 许可证、BSD 许可证、Mozilla 公共许可证等。MIT 许可证是一种宽松的许可证，几乎没有限制，允许用户自由地使用、复制、修改、合并、出版、分发、再许可和销售软件；GPL 许可证要求在所有衍生作品中使用相同的许可证，即如果用户使用了 GPL 许可证下的代码或项目，则用户的整个项目也必须遵循 GPL 许可证；Apache 许可证与 MIT 许可证类似，但需要声明出处、保留版权和贡献清单等要求；BSD 许可证也类似于 MIT 许可证，但增加了需要在分发时提供原始许可证和版权声明等一些其他限制；Mozilla 公共许可证与 GPL 许可证类似，更适合于涉及网络浏览器和其他公共网络服务等方面的项目。开源许可证各有不同的特点和适用场景，贡献者需要根据具体需求和项目的性质选择合适的开源许可证。

主题 2 的高频单词商业、模式、公司等符合开源软件商业模式的研究，被标注为“商业模式”。开源软件商业模式也得到了学者的广泛关注，包括选择不同商业模式的因素[50]、如何设计商业模式[51]以及系统地对商业模式的种类进行总结等[52]，商业模式的选择和设计影响着开源项目的成功。常见的几种商业模式有以下几种：第一种是开放核心，即软件核心代码部分开源，非核心部分闭源，对软件的部分插件或者运行时所需要的素材收费，从而通过提供差异化的商业产品来为客户提供服务；第二种是支持和咨询服务收费，公司基于开源软件提供技术支持、培训和咨询服务，利用自己对开源软件的专业知识和经验，为企业提供定制化的支持和解决方案，并提供付费的服务合同；第三种是延迟开源模式，即新版本闭源，旧版本开源的模式，当公司研发出更新的商业版本之后，原来的商业版本就会被开源出来；第四种是双重许可模式，公司采用双重许可模式，即将开源软件以开源许可证发布，但同时也提供商业许可证，开源许可证使软件在开放源代码下免费使用和修改，而商业许可证则允许客户在某些条件下获得额外的权益和功能。第五种是捐赠和赞助模式，公司通过接受捐赠和赞助来支持开源项目的开发和维护，他们在开源社区中建立信誉和影响力，并通过向企业、组织或个人寻求资金支持来确保项目的可持续性。开源与商业化相辅相成，Red Hat 作为一家领先的开源技术公司，通过提供企业级支持和服务以及开源云计算、容器化、存储和中间件解决方案等已经取得了巨大的成功，Red Hat 的 Linux 发行版被广泛用于企业级应用，因此为客户提供技术支持非常重要；此外，基于文档存储的 NoSQL 数据库 MongoDB、提供日志分析和数据可视化的开源软件公司 Elastic 以及基于容器技术的开源平台 Docker 都是成功的开源商业公司。

主题 3 中治理、模式、网络、生态系统等高频词贴合开源软件社区治理的研究内容，被标注为“开源治理”。开源治理也是开源领域的热门话题。众所周知，诸如开源社区等虚拟社区的组织治理方式与传统组织不同[53]，开源项目所有者必须求助于其他治理机制，而不是那些向开发人员付费的公司所提供的治理机制[54]。Vishal Midha 等[55]提出开源项目治理的二维分类，即参与管理和责任管理，并展示了两个治理维度对开源软件维护结果的影响。Saerom Lee 等[56]探讨了将开发人员分配到组织内多个项目的有效治理策略，以促进协作软件开发和简化协调。早期关于治理的研究大多聚焦于开源社区，随着越来越多的公司参与开源，学者开始关注对企业开源治理的研究[57][58]，探究公司如何在开源和闭源之间做出权衡[59]以及如何分配员工参与开源的时间[60]等。

主题 6 的高频单词社区、公司、开发者则体现了对公司内部员工参与开源项目的研究，被标注为“企业参与”。开源软件在过去几十年中得到了广泛的应用和发展，并吸引了许多企业参与其中。企业参与开源软件领域的活动能够获得最新的技术发展动态[61]，与其他开发者共享经验和知识，推动技术创新[62]；可以减少开发成本[59]，提升其品牌知名度和声誉，树立技术领导力和社区贡献形象。企业可以通过提供与开源软件相关的技术支持和培训服务，培养一个由同行生产者组成的开源社区[61]，鼓励员工向开源项目贡献代码、功能和修复漏洞，组织和赞助开源软件相关的活动等来为开源项目的发展做出贡献[66]。

4.2 热点主题识别

主题强度反映在一段时间内一个主题在文档集中的相对重要性或突出程度。热点主题表现为在文档集中频繁出现的主题，即在特定时间段内主题强度值相对较高的主题，主题强度越大越有可能被认为是热点主题[63]。通过 LDA 主题建模后得到的文档-主题矩阵可以计算每一年的主题强度，文档-主题矩阵中包含了每一篇文档

在每一个主题上的概率值，将每一个主题在某一年内所有文档上的概率值相加并求平均，就可以得到该主题在该年份上的主题强度数值，主题强度计算方式如公式 3 所示。

$$S_k^t = \frac{\sum_{d=1}^{M}\theta_{d,k}}{M} \qquad \text{公式 3}$$

其中 M 表示 t 年份文档总数，如果计算总体主题强度，则 M 表示所有文档总数。$\theta_{d,k}$表示第 d 篇文档上第 k 个主题的概率值。在进行热点主题识别时，热点主题的阈值定义为所有主题强度的平均值，超过主题强度阈值的主题即为热点主题。具体结果如图 8 所示。



图 8 主题强度分布图

可以看出，贡献动机、商业模式、协作模式为开源软件领域研究的热点主题。首先，开源软件参与者的参与动机一直以来都是开源软件研究者感兴趣的话题，在开源运动中，贡献者根据自己的能力、兴趣爱好，选择感兴趣的项目并为项目做出贡献，开源软件的参与者往往包括个人、企业、基金会、政府等[64]，其中，个体的参与最受研究者关注，其次是企业。参与开源软件开发的个体不仅包括自由职业者，还包括企业的员工，他们出于不同的动机对软件项目做出贡献。其次，在开源软件开发社区中，由于不同的贡献者位于全球各地，且开发过程中涉及到复杂的技术、知识和沟通问题，因此，对开发人员协作模式研究的热度也相对较高。此外，学术界对于开源软件商业模式的研究也是一个相对活跃的领域，并且开源软件在企业和社会之间普及的过程中，越来越多的学者开始关注开源软件如何实现商业化运营，由此也带动了学术界对于开源社区治理、公司员工参与开源的模式的研究。相比而言，学者对于开源软件项目许可证选择的研究相对较少。

国内外学术界对开源软件研究热度的不断增强，促进了学科交叉融合和知识共享。但从图 9 可以看出，中国对于开源共享的研究与国外仍有较大差距，要想为中国开源共享的发展提供理论支撑和实践经验，理论研究还有待加强。

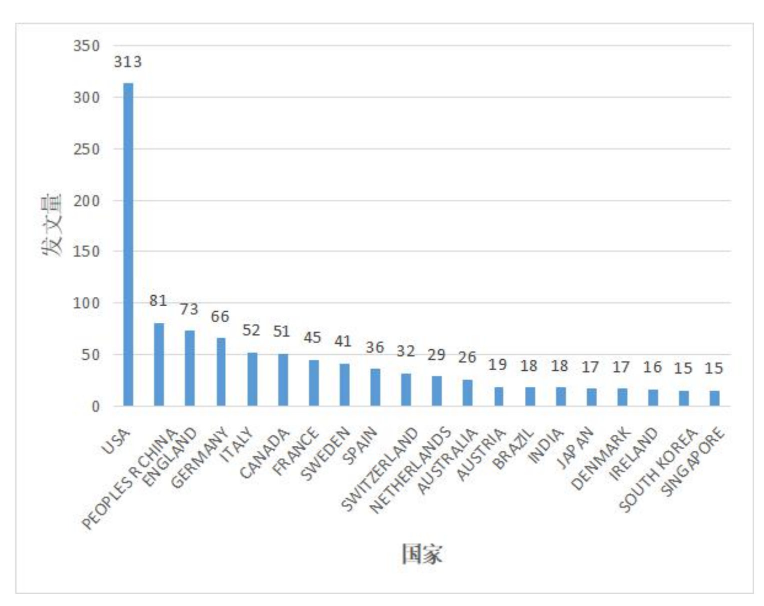


图 9 开源软件研究的国家分布

4.3 主题演化分析

开源软件的演化过程中，主题强度的变化对于了解主题的热度和重要性具有重要意义。主题强度演化可以帮助我们揭示开源软件社区中不同主题的变化趋势及影响力。在每一年内根据公式 3 计算当年每一个主题的主题强度，根据计算得到的主题强度数据，进行主题强度的演化分析。由于各个主题强度变化的波动性较强，难以看出明显的趋势，本文对原始数据做了三期移动平均处理，减少数据波动带来的影响，并依照先前学者将主题强度的演化分为上升型、下降型和稳定型[65]，得出的演化结果如图 9-图 11 所示。

从图 10 可以看出，商业模式、开源治理、企业参与主题有波动上升的趋势。商业模式主题、开源治理主题、企业参与主题分别研究软件开源的商业模式、开源治理问题以及公司员工参与开源项目的问题。近年来在软件开源领域，对于软件开源的商业模式、开源治理问题以及公司选择员工参与开源项目的研究引起了广泛的关注和研究。由于开源软件已经在许多行业和领域中得到广泛应用，学者对如何在开源软件项目中创造商业价值、构建可持续的商业模式等问题的研究兴趣增加，因此开源软件在商业化方面的研究逐渐增多，这些研究对于帮助企业更好地利用开源软件、促进商业创新具有重要意义。

另一方面，对于开源软件社区治理问题的研究也越来越受关注。一个优秀的开源项目，更重要的是维护一个开源社区，开源社区包含了一套完整的项目管理流程，它包括开放源代码、社区写作流程、项目质量管理为一体的系统管理。开源软件社区的治理是保证开源软件项目顺利进行的关键因素，学者通过研究社区的组织结构、决策过程、资源分配等方面的问题，试图提出更好的开源治理模式和方法，促进开源社区可持续发展。一般一个大型的开源社区有以下五种角色[66]：（1）开源领导者（Leader）：领导者承担了带领项目发展的责任，一般拥有项目事务的决策权。（2）开源维护者（Maintainer）：维护者承担了项目日常维护工作，一般拥有项目事务的管理权，开源维护者是项目中的主要管理者，会帮助开源领导者分担项目管理事务。（3）开源提交者（Committer）：提交者负责对项目提交项目成果（一般指源代码提交），并参与项目事务的处理，开源提交者是可以直接提交代码到主干的人，在项目维护的模块中发挥重要作用。（4）开源贡献者（Contributor）：贡献者通过多种方式为项目做贡献（如创建问题、打开讨论、回答讨论、提议拉取请求、提交拉取请求等），贡献者可以通过提交 PR、提交 Issues、解决 Issues、帮助项目写文档、邮件反馈、社区分享、社区答疑、宣传推广等方式，为开源项目贡献自己的一份力量。（5）开源使用者（User）：使用者是开源项目的使用者，一般会围绕项目进行技术讨论和意见反馈，开源使用者作为社区成员，他们最有价值的部分是提出需求、报告缺陷、提出建议。虽然目前中国开源社区数量很多，但开源社区的运营和治理能力大部分还处于比较初级的阶段，形式上具备国外开源社区的治理架构，但还没有真正发挥出开源开放和协作的效应，社区贡献主要还是来源于项目的发起方，开源社区的治理问题有待进一步研究。

此外，公司员工参与开源项目的相关研究也逐渐增多，越来越多的公司意识到参与开源项目可以为他们带

来多方面的好处，包括技术增长、声誉提升、人才招聘等。研究人员开始探索公司如何选择合适的员工参与开源项目，如何管理员工的开源参与以实现最大化的利益。

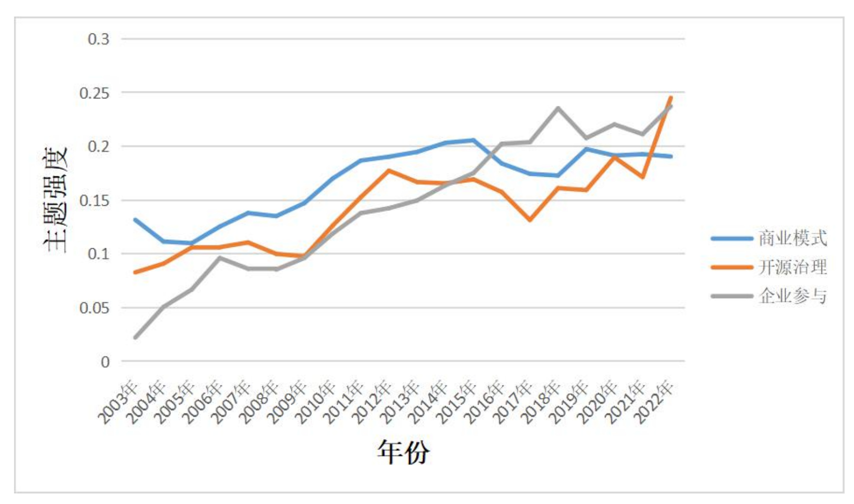


图 10 上升型主题强度曲线

从图 11 可以看出，贡献动机主题和协作模式主题研究强度相对下降，贡献动机主题和协作模式主题分别研究贡献者参与开源项目的动机和贡献者彼此之间的协作模式。在软件开源领域，早期的研究主要集中在分析贡献者参与开源项目的动机和贡献者之间的协作模式，这些研究可以帮助我们更好地理解开源软件社区的运作机制，促进开源软件的发展和持续改进。随着开源软件社区的不断发展和成熟，越来越多的研究开始涉及到开源软件的商业化、治理问题、安全性等更加深入的话题。因此，对于贡献者参与开源项目的动机和贡献者之间的协作模式的研究相对减少，主题强度下降。此外，随着开源软件的普及，越来越多的人开始参与到开源软件社区中来，使得开源软件社区的规模不断扩大，贡献者之间的互动和合作变得更加复杂和多样化。因此，简单地对贡献者参与开源项目的动机和贡献者之间的协作模式进行研究已经不能很好地解决开源软件社区面临的各种问题，需要更加综合和深入的研究方法和视角。

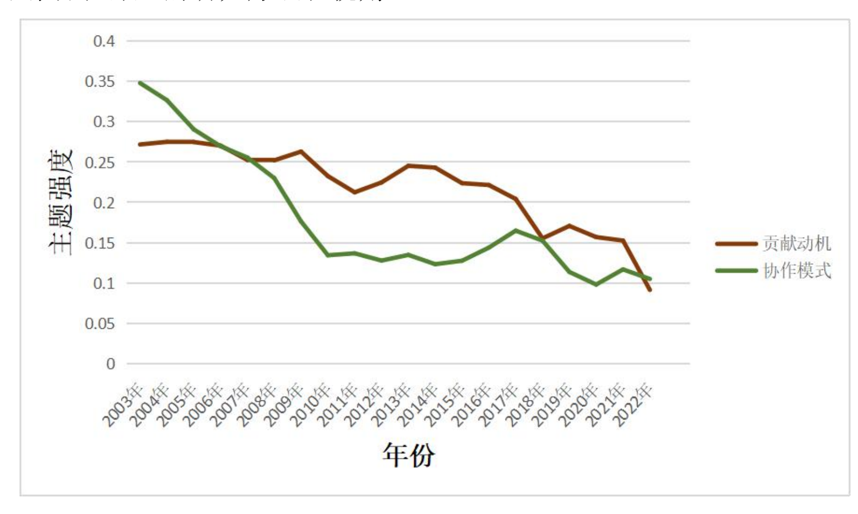


图 11 下降型主题强度曲线

图 12 显示开源协议研究强度处于一个相对稳定的状态。开源协议主题主要研究开源许可证的选择，也涉及到项目成功因素的探讨。作为开源软件研究领域的基础问题，开源软件许可证选择和开源项目成功因素的研究总体上一直保持稳定的状态。开发者在免费获得开源软件源代码的同时，仍需遵守开源协议，不可随意使用。为了使开源软件更合理、规范的使用，保护开源软件的知识产权，OSI 已经认证通过了 80 多个开源软件许可证，用法律的手段为开源软件的使用、修改、复制、分发进行规范。开源软件许可证的选择涉及到权利和义务的平衡，对于项目的法律合规性和商业可持续性具有重要影响。项目成功因素的研究可以帮助贡献者和组织了解如何提高项目的质量、吸引贡献者、增加用户参与等方面的问题。学者对开源许可证的选择和项目成功因素

的研究，可以为贡献者、组织和学术界提供有益的指导和深入的理论洞察，促进软件开源的可持续发展。

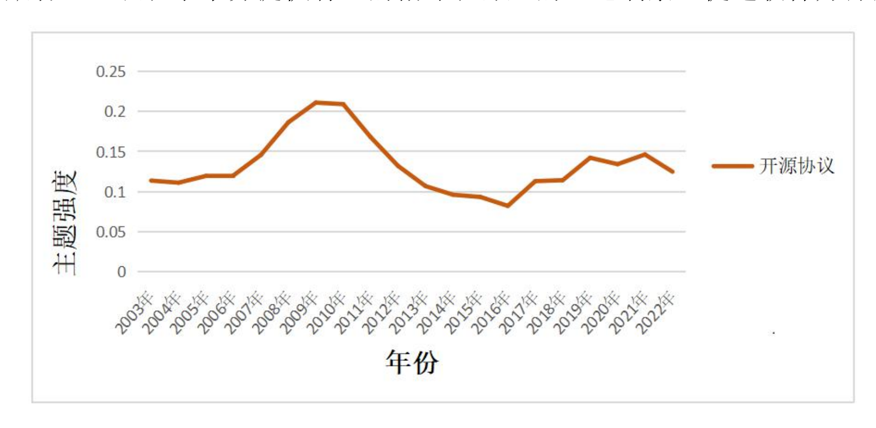


图 12 稳定型主题强度曲线

5 结论及建议

5.1 结论

软件开源生产模式对软件创新产业格局产生了深远的影响，加速了知识创新的过程。软件开源领域的研究主要覆盖了开源生态六个方面的主题，从整体研究的主题强度上看，贡献动机、商业模式和协作模式三个主题的强度超过主题强度的阈值，研究强度高，属于软件开源研究领域的热点主题。从主题强度演化趋势上看，软件开源研究与开源生态的演化具有一致性，即呈现从最初简单的个人维度关注开源动机的自发性和自治性，逐渐向复杂组织维度的企业参与和政府宏观治理演变的发展规律。针对国外领先的开源社区如 github，研究者对于开源商业模式、开源治理、企业参与的研究处于波动上升的趋势，研究热度增加；对于开源协议主题的研究始终处于相对稳定状态，而对研究贡献者贡献动机和贡献者协作模式方面的研究热度绝对值稳定，但相对比重下降。

5.2 建议

软件开源研究主题发展演化规律某种程度上代表了软件开源运动所关注焦点问题的变化规律。目前我国正虽是软件生产大国，但在基础应用软件和专业生产软件等领域仍存在明显短板，构建以社会协同为特点的软件创新开源生态是必不可少的发展路径。良好的开源生态需要生态中各个参与者共同努力，建议学者们针对中国情景开展研究，为构建开源创新生态服务中国软件产业做出贡献。

个人层面，建议开展中国情景下的开源贡献者参与动机研究，研究并设计能够有效鼓励个人积极在开源社区贡献的开源创新机制及文化氛围；企业层面，研究企业参与开源的动机及机制设计，为企业开源的战略决策提供理论指导；平台及组织层面，研究设计适合中国情景的开源社区治理方案等。同时，研究基于开源社区的创新创业案例和模式，鼓励基于开源的创业，为全球提供更多中国实践与智慧。

参考文献：

[1] 周涛,王超.开源软件社区用户知识贡献行为研究[J].科研管理,2020,41(02):202-209.
[2] 中国人大网.《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》[EB/OL]. http://www.npc.gov.cn/，2021-03-13.
[3] 中华人民共和国工业和信息化部.《“十四五”软件和信息技术服务业发展规划》[EB/OL]. http://www.npc.gov.cn/，2021-11-30.
[4] 中国政府网.《“十四五”数字经济发展规划》[EB/OL]. https://www.gov.cn/.2022-1-12.
[5] 卢国强,黄微,孙悦等.基于舆情客体与本体剥离的重大突发事件网络舆情本体演化强度研究[J].图书情报工作,2023,67(05):119-129.DOI:10.13266/j.issn.0252-3116.2023.05.011.
[6] 马晓悦,薛鹏珍,陈忆金等.社交媒体危机主题演化模型构建与趋势分析[J].图书情报工作,2021,65(13):77-86. DOI:10.13266/j.issn.0252-3116.2021.13.008.
[7] Huangfu L, Mo Y, Zhang P, et al. COVID-19 vaccine tweets after vaccine rollout: sentiment－based topic modeling[J]. Journal of medical Internet research, 2022, 24(2): e31726.

[8] Qian Y, Liu Y, Sheng Q Z. Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence[J]. Journal of Informetrics, 2020, 14(3): 101047.
[9] 张玲,恽诚涛,尹思力等.我国科研诚信政策与文献主题演化对比分析[J].现代情报,2023,43(06):108-120.
[10] 李秀霞,程结晶,韩霞.发文趋势与引文趋势融合的学科研究主题优先级排序——以我国情报学学科主题为例[J].图书情报工作,2019,63(11):88-95.
[11] 唐果媛.基于共词分析法的学科主题演化研究方法的构建[J].图书情报工作,2017,61(23):100-107.
[12] 丁晟春,刘笑迎,李真.融合评论影响力的网络舆情热点主题演化研究[J].现代情报,2021,41(08):87-97.
[13] Huang C, Yang C, Wang S, et al. Evolution of topics in education research: A systematic review using bibliometric analysis[J]. Educational Review, 2020, 72(3): 281-297.
[14] 邢晓昭,任亮,雷孝平等.基于专利主题演化的颠覆性技术识别研究——以类脑智能领域为例[J].情报科学,2023,41(03):81-88.
[15] Vahidzadeh R, Bertanza G, Sbaffoni S, et al. Regional industrial symbiosis: A review based on social network analysis[J]. Journal of Cleaner Production, 2021, 280: 124054.
[16] 曾子明,陈思语.基于 LDA 与 Bert-BiLSTM-Attention 模型的突发公共卫生事件网络舆情演化分析[J/OL].情报理论与实践:1-13[2023-05-22].http://kns.cnki.net/kcms/detail/11.1762.G3.20230412.0904.002.html
[17] 周健,张杰,屈冉等.基于 LDA 的国内外区块链主题挖掘与演化分析[J].情报杂志,2021,40(09):161-169.
[18] Liu J, Nie H, Li S, et al. Tracing the pace of COVID-19 research: topic modeling and evolution[J]. Big Data Research, 2021, 25: 100236.
[19] 张柳,王慧,相甍甍.基于 LDA 的突发事件应急管理主题热度与演化分析[J/OL].情报科学:1-20 [2023-05-22].http://kns.cnki.net/kcms/detail/22.1264.g2.20230509.1445.018.html
[20] 陈光沛,魏江,李拓宇.开源社区：研究脉络、知识框架和研究展望[J].外国经济与管理,2021,43(02):84-102.
[21] Wang Z, Perry D E, Xu X. Characterizing individualized coding contributions of OSS developers from topic perspective[J]. International Journal of Software Engineering and Knowledge Engineering, 2017, 27(01): 91-124.
[22] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
[23] Zhou, H., Yu, H. & Hu, R. Topic evolution based on the probabilistic topic model: a review. Front. Comput. Sci. 11, 786 – 802 (2017).
[24] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 178-185.
[25] 夏萌萌,汝绪伟,张红军.基于 LDA 模型的产业创新生态系统研究主题演化分析[J].中国高校科技,2022,No.409(09):41-46.
[26] 陈琦,张君冬,郑婉婷等.基于 LDA 模型的中医药人工智能领域主题演化分析[J].世界科学技术-中医药现代化,2022,24(09):3315-3324.
[27] 贺亮,李芳.基于话题模型的科技文献话题发现和趋势分析[J].中文信息学报,2012,26(02):109-115.
[28] 冉从敬,李旺.基于 LDA 的企业竞争对手识别模型构建——以蔚来汽车有限公司为例[J/OL].情报理论与实践:1-11[2023-05-25].http://kns.cnki.net/kcms/detail/11.1762.G3.20230410.1717.006.html.
[29] 谭春辉, 熊梦媛. 基于 LDA 模型的国内外数据挖掘研究热点主题演化对比分析[J]. 情报科学, 2021, 39(4): 174-185.
[30] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics[C]//Proceedings of the workshop on interactive language learning, visualization, and interfaces. 2014: 63-70.
[31] Lerner J, Tirole J. Some simple economics of open source[J]. The journal of industrial economics, 2002, 50(2): 197-234.
[32] Alexander Hars S O. Working for free? Motivations for participating in open-source projects[J]. International journal of electronic commerce, 2002, 6(3): 25-39.
[33] Ye Y, Kishida K. Toward an understanding of the motivation of open source software developers[C]//25th International Conference on Software Engineering, 2003. Proceedings. IEEE, 2003: 419-429.
[34] Eric von Hippel,Georg von Krogh.Open source software and the“private-collective” innovation model: Issues for organization science[J].Organization Science，2003，14(2):209-223.
[35] Oreg,S and Nov,O.Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values[J].Computers in Human Behavior,2008,24(5):2055-2073.
[36] Krogh G V , Haefliger S , Spaeth S , et al. Carrots and Rainbows: Motivation and Social Practice in Open Source Software Development[J]. Mis Quarterly, 2012, 36(2):649-676

[37] Moqri,M.;Mei,X.;Qiu,L.;and Bandyopadhyay,S.Effect of “following” on contributions to open source communities[J].Journal of Management Information Systems,2018,35(4):1188–1217.
[38] Raymond E. The cathedral and the bazaar[J]. Knowledge, Technology & Policy, 1999, 12(3): 23-49.
[39] Hong Q, Kim S, Cheung S C, et al. Understanding a developer social network and its evolution[C]//2011 27th IEEE international conference on software maintenance (ICSM). IEEE, 2011: 323-332.
[40] Hahn J, Moon J Y, Zhang C. Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties[J]. Information Systems Research, 2008, 19(3): 369-391.
[41] Stewart K J, Ammeter A P, Maruping L M. Impacts of license choice and organizational sponsorship on user interest and development activity in open source software projects[J]. Information Systems Research, 2006, 17(2): 126-144.
[42] Fang Y, Neufeld D. Understanding sustained participation in open source software projects[J]. Journal of Management Information Systems, 2009, 25(4): 9-50.
[43] Gamalielsson J, Lundell B. Sustainability of Open Source software communities beyond a fork: How and why has the LibreOffice project evolved?[J]. Journal of Systems and Software, 2014, 89: 128-145.
[44] Sha Z, Petrov A, Tian Y, et al. Analyzing the Robustness of Open Source Software Ecosystems to the Loss of Contributors: A Case Study[J]. Available at SSRN 4082801.
[45] SL Daniel L M, Maruping L, Cataldo M, et al. The impact of ideology misfit on open source software communities and companies[J]. Management information systems quarterly, 2018, 42(4).
[46] Maruping L M, Daniel S L, Cataldo M. Developer centrality and the impact of value congruence and incongruence on commitment and code contribution activity in open source software communities[J]. MIS Quarterly, 2019, 43(3): 951-976.
[47] Tang T Y, Fang E E, Qualls W J. More Is Not Necessarily Better: An Absorptive Capacity Perspective on Network Effects in Open Source Software Development Communities[J]. MIS Quarterly, 2020, 44(4).
[48] Gangadharan G R, D’Andrea V, De Paoli S, et al. Managing license compliance in free and open source software development[J]. Information Systems Frontiers, 2012, 14: 143-154.
[49] Sen R, Subramaniam C, Nelson M L. Open source software licenses: Strong-copyleft, non-copyleft, or somewhere in between?[J]. Decision support systems, 2011, 52(1): 199-206.
[50] Perr J, Appleyard M M, Patrick P. Open for business: emerging business models in open source software[J]. International Journal of Technology Management, 2010, 52(3/4): 432-456.
[51] Belenzon S, Schankerman M. Motivation and sorting of human capital in open innovation[J]. Strategic Management Journal, 2015, 36(6): 795-820.
[52] Shahrivar S, Elahi S, Hassanzadeh A, et al. A business model for commercial open source software: A systematic literature review[J]. Information and Software Technology, 2018, 103: 202-214.
[53] Ferraz I N, Santos C D. Transformation of free and open source software development projects: governance between the cathedral and bazaar[J]. Revista de Administração de Empresas, 2022, 62.
[54] Von Krogh G, Von Hippel E. The promise of research on open source software[J]. Management science, 2006, 52(7): 975-983.
[55] Midha V, Bhattacherjee A. Governance practices and software maintenance: A study of open source projects[J]. Decision Support Systems, 2012, 54(1): 23-32.
[56] Lee S, Baek H, Jahng J. Governance strategies for open collaboration: Focusing on resource allocation in open source software development organizations. International Journal of Information Management. 2017. pp. 431–437.
[57] Schaarschmidt M, Walsh G, von Kortzfleisch H F O. How do firms influence open source software communities? A framework and empirical analysis of different governance modes[J]. Information and Organization, 2015, 25(2): 99-114.
[58] Harutyunyan N, Riehle D. Getting started with corporate open source governance: A case study evaluation of industry best practices[J]. 2021.
[59] Wang Y, Chen Y, Koo B. Open to your rival: Competition between open source and proprietary software under indirect network effects[J]. Journal of Management Information Systems, 2020, 37(4): 1128-1154.
[60] Mehra A, Mookerjee V. Human capital development for programmers using open source software[J]. MIS quarterly, 2012: 107-122.
[61] Rolandsson B, Bergquist M, Ljungberg J. Open source in the firm: Opening up professional practices of software development[J]. Research Policy, 2011, 40(4): 576-587.
[62] Dahlander L, Magnusson M. How do firms make use of open source communities?[J]. Long range planning, 2008, 41(6): 629-649.

[63] 林丽丽,马秀峰.基于 LDA 模型的国内图书情报学研究主题发现及演化分析[J].情报科学,2019,37(12):87-92. DOI:10.13833/j.issn.1007-7634.2019.12.013.
[64] 范小青.我国开源运动参与者的参与动机研究[J].教育传媒研究,2019(01):18-25.
[65] 林丽丽,马秀峰.基于 LDA 模型的国内图书情报学研究主题发现及演化分析[J].情报科学,2019,37(12):87-92.
[66] Eghbal N. Working in public: the making and maintenance of open source software[M]. San Francisco: Stripe Press, 2020.

作者贡献说明：

董平军：确定选题，指导论文写作，修改和审定论文；
高翔菲：文献调研与资料收集，数据处理，论文撰写。

Topic Mining and Evolution Analysis of Software Open Source Research

Dong Pingjun　Gao Xiangfei
Rising Sun School of Business Administration, Donghua University,　Shanghai　200051

Abstract：[Purpose/Significance] Software open-source is an important production organization and collaborative innovation movement in socialized software production. By identifying and analyzing the themes and evolution of software open-source related research at home and abroad, this study explores the phased hotspots and trend changes in the field of software open-source research, and provides research direction for scholars with the main purpose of promoting further optimization and development of software open-source innovation in China. [Method/Process] This paper uses the software open source literature retrieved from the Web of Science database from 2001 to May 10, 2023 as the corpus, uses the Perplexity index to determine the number of topics, trains the LDA topic recognition model to obtain the topic word distribution and document topic distribution, identifies topics according to the topic word distribution, calculates the topic intensity according to the document topic distribution, and then identifies hot topics and summarizes the evolution path. [Result/Conclusion] The results of topic identification indicate that there are six important themes in the field of software open source research, namely contribution motivation, business model, open source governance, collaboration model, open source protocol, and enterprise participation; From the perspective of theme evolution, software open-source has shown relatively high research enthusiasm in business models, open source governance, and enterprise participation themes in recent years. The research trend of open source protocols is relatively stable, and although the research enthusiasm for contribution motivation and collaboration models is relatively declining, it has always maintained a high level of attention from beginning to end. The research on software open source presents a development pattern from the individual dimension of spontaneous and autonomous attention to open source motivation to the organizational dimension of enterprise and government participation. It is recommended that scholars pay attention to various thematic studies on the open source ecosystem in the Chinese context, in order to provide theoretical support for the healthy development of the open source ecosystem in China.

Keywords：open source software　topic identification　topic evolution　LDA model